3-27-2015

Twitter Event Detection

David Kimmey
Indiana University - Purdue University Fort Wayne

Twitter Event Detection

David L. Kimmey

Department of Computer Science - Purdue University Fort Wayne * Fort Wayne, Indiana 46805

Faculty Sponsor: Jin Soung Yoo

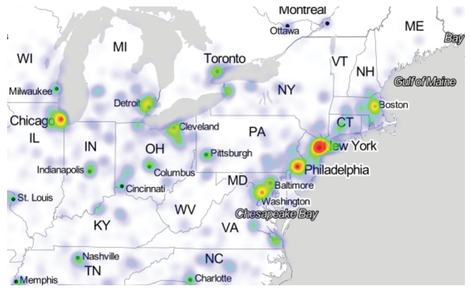


Figure 1: Heatmap of Influenza Tweets, March 16, 2015
Number of Tweets per area (i.e. density) of tweets related to influenza
http://nowtrending.hhs.gov/heatmap?taxonomy=influenza

1. Introduction

What methods are available for Twitter event detection? Predicting the present (i.e. "contemporaneous forecasting" or "nowcasting") is a hot topic of interest to corporations, central banks, and other government agencies. Furthermore, Twitter data is becoming a popular source of social trajectory data for "nowcasting" (e.g. geo-economic events, the discovery of unusual social events, geographic disease/influenza trends (Figure 1), and social questions). However, the defining and detecting of events (i.e. a thing that happens, especially one of importance) is a non-trivial task.

2. Goal

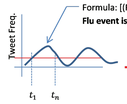
Given: a set of raw (i.e. unprocessed) Twitter data
Output: find a flu event, and the list of Tweets that occur during the event for use in future data mining work (e.g. descriptive and predictive models of the flu event).

3. Methods

Data: raw flu Tweets--unprocessed

Tweet Frequency Bursts Method

Frequency Bursts [1]:
Formula: [(Peak frequency rate) / (mean frequency rate)] > 1
Flu event is detected by the Bursts (i.e. spikes)

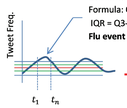


Flu event is defined as the Tweets between time t_1 and t_n

List of Tweets during flu event
Tweet 1: buy a set of tiers
Tweet 2: cat in the hat
Tweet 3: i am sick today
...
Tweet n: going home today

Tweet Deviation / Anomaly Detection Method

Deviation / Anomaly Detection [2]:
Formula: Q1 = lower quantile; Q3 = upper quantile;
IQR = Q3-Q1; Outlier = Q1-1.5*IQR or Q3+1.5*IQR
Flu event is detected by the Outliers

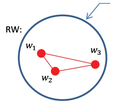


Flu event is defined as the Tweets between time t_1 and t_n

List of Tweets during flu event
Tweet 1: need a doctor
Tweet 2: i feel ill
Tweet 3: i am sick today
...
Tweet n: the car is sick

Tweet Probabilistic Soft Logic Method

Probabilistic Soft Logic (PSL) [3]:
Formula:
RW: $\text{trigger\_event}_1(w_1) \wedge \text{trigger\_event}_2(w_2) \Rightarrow \text{final\_event}(w_3)$
where RW=rule weight, and w_n = soft truth value



Flu event is defined as the Tweets from the PSL formula

List of Tweets during flue event
Tweet 1: need medicine
Tweet 2: i feel sick
Tweet 3: do you have flu
...
Tweet n: do you know a Dr.

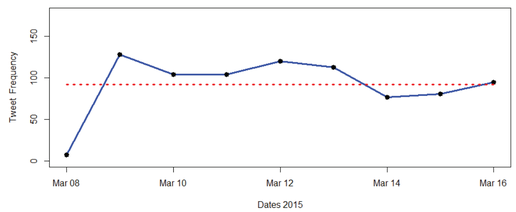
Future Work using the Event Detection Methods:

Use the Tweets found through these Methods: Frequency Bursts, Deviation / Anomaly Detection, and Probabilistic Soft Logic → Run Tweets through Sentiment Analysis to generate Tweet labels: Positive, Negative, and Neutral → Use labels in a Decision Tree Classification to build a descriptive and predictive model of the current flu event trend

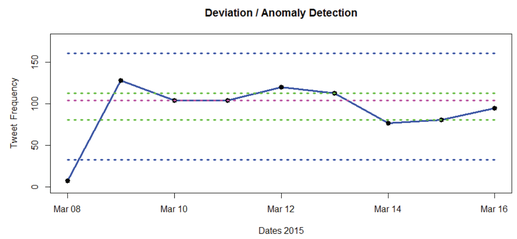
4. Results

Tweet Frequency Bursts: flu event detected in Tweets from March 9th to March13th 2015

Twitter Event Detection Method output



Tweet Deviation / Anomaly Detection: no flu event detected in Tweets (Tweet freq. is not above top blue line)

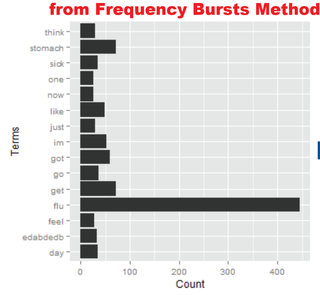


Tweet Probabilistic Soft Logic: flu event detected in Tweets by PSL Rules from March 8th to March 16th 2015

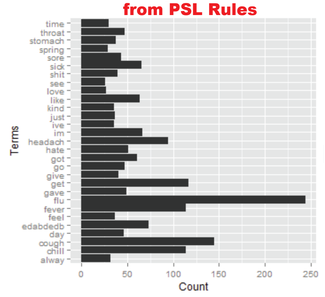
Probabilistic Soft Logic Rules

0.9782340:cough(0.35522180)^cough(0.35522180)=>flu(0.2895564)
0.9778025:headaches(0.32756233)^stuffy+nose(0.09833795)=>flu(0.5740997)
0.9652203:cough(0.35835095)^chills(0.34954193)=>flu(0.2921071)
0.9596413:headaches(0.20619006)^chills(0.43243243)=>flu(0.3613775)
0.9253672:cough(0.35835095)^chills(0.34954193)=>flu(0.2921071)
0.9227623:chills(0.47623620)^sore+throat(0.12578012)=>flu(0.3979837)
...
0.7254383:cough(0.3552218)^cough(0.3552218)=>flu(0.2895564)

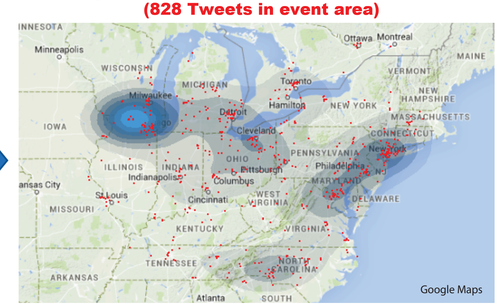
Twitter Terms used in flu event and term frequency counts > 25 from Frequency Bursts Method



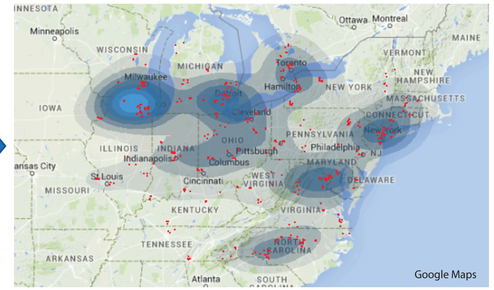
Twitter Terms used in flu event and term frequency counts > 25 from PSL Rules



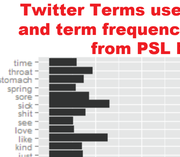
Freq. Bursts Density Map of flu event Tweets (828 Tweets in event area)



PSL Density Map of flu event Tweets (828 Tweets in event area)



5. Summary



Text Cloud of PSL results

- The dark gray and blue density areas of the PSL density map align closer to the red, yellow, and green density areas of the U.S. Department of Health's Heatmap (Figure 1). Therefore, the PSL method provides better results than the Frequency Bursts and the Deviation/Anomaly detection methods.
- The PSL method produced the largest corpus of tweets with more words and higher frequency counts per word. The larger corpus will be an advantage in future data mining work.
- The Deviation/Anomaly Detection method was not able to find this event because of the event's low frequency of tweets.

6. Conclusion

Overall, the Probabilistic Soft Logic (PSL) method identified more of the event area than the Frequency Bursts and Deviation/Anomaly detection methods. In addition, PSL seems to have the advantage of finding events in low Tweet frequency environments, and PSL is not dependent on the selection of the time series window (i.e. tweets per hour vs tweets per day). Future work will be needed with larger data sets, and to see if the tweets from the PSL method will produce an effective model after data mining.

References

[1] J. Weng, Y. Yao, E. Leonardi, and B. Lee, Event Detection in Twitter, ICWSM, aaai.org, 2011
[2] R. Lee, S. Wakamiya, and K. Sumiya, Discovery of unusual regional social activities using geo-tagged microblogs, Springer Science+Business Media, LLC, 2011
[3] R. Santos, S. Shah, F. Chen, A. Boedihardjo, C. Lu, and N. Ramakrishnan, Forecasting Location-based Events with Spatio-temporal Storytelling, ACM SIGSPATIAL LBSN '14, November 4, 2014